Tools for Research Projects

March 2018

Stats4SD

This document was produced in collaboration with Statistics for Sustainable Development and is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Introduction

In this document we look at some of the tools that a researcher might use during the lifetime of a project. We have included a brief paragraph about each tool mentioned and have included web links where researchers can find further information. This document is intended to be a "live" document which should be regularly updated as new tools become available; web links should also be checked periodically to ensure they are still valid.

This document is not intended to be a list of recommendations; we do not promote any particular software solution but simply wish to share information about tools that we have used and have found useful for research projects.

Data Collection Tools

In this section we look at a few tools that can be used for data collection and data entry. Some of these are specifically used for data collection on hand-held devices; e.g. ODK; while others, such as CSPro can also be used for entering data from paper questionnaires.

Open Data Kit (ODK)

ODK is a free and open-source set of tools for building data collection forms for Android devices. Forms are generally built using the XLSForm standard, which was created to help simplify the authoring of forms using Excel. It is relatively easy to get started on using ODK and as you become more proficient you can build up to quite complex forms including skip logic, data checks, and repeating groups. For information on writing data entry forms for ODK, see http://xlsform.org/ .

Once the form is complete it is converted to a format that can be installed on handheld devices. ODK Collect is then used on the device to collect the data during the interview. Once the data have been collected and checked, they can be uploaded to an aggregate server from where they can be analysed or exported. This process allows for monitoring of data collection, and progress reports can be quickly and easily produced during the fieldwork.

ODK is typically used with a server, which acts as the central point for uploading your forms and aggregating collected data from multiple devices and users. There are many server options available to projects, including third-party services (both free / open and commercial) and self-hosting options.

Your organisation's IT / Data Management supporters will likely have recommended services to use, and your organisation may even have their own ODK Aggregate server that you can use for your project. We recommend discussing your requirements within your organisation to learn if such services are available.

For further information on ODK go to https://opendatakit.org/

CSPro

The Census and Survey Processing System (CSPro) is a public domain software package for entering and managing census and survey data. It is developed and supported by the US Census Bureau.

CSPro is designed to be as user-friendly as possible yet is powerful enough to handle complex applications. The software runs on Windows PCs but there is also an CSEntry Android App which works in collaboration with the desktop version of CSPro. Data entry applications would be created under Windows and compiled versions would be transferred to the Android device for data collection. Data files can then be uploaded to Dropbox or to an FTP server while in the field.

Alternatively, data from paper questionnaires can be entered directly into the Windows application using forms that can be designed to resemble the questionnaire.

For further information on CSPro go to https://www.census.gov/data/software/cspro.Overview.html

Surveybe

Surveybe is a computer assisted personal interviewing (CAPI) software suite. The suite comprises two elements: the Designer and the Implementer.

The Designer is used to build, configure and update a questionnaire, including screen structure, questions, rosters and validations. The Implementer is then used to display the questionnaire on the device and collect the data.

Note: Surveybe is not a free product; you would need to buy a license for the Designer. However, when you buy a Designer license, the Implementer is free. Thus, with a single Designer license you can do all your designing on a single PC but use the Implementer on any number of devices for data collection.

For further information on Surveybe, including pricing, go to http://surveybe.com/

KoBoToolbox

KoBoToolbox is a suite of tools that uses ODK for field data collection. The main service is the Kobotools server, which fulfils the role of the ODK Aggregate server. This service is freely provided, as it is funded by the Harvard Humanitarian Initiative and a group of other international organisations.

Kobotools also has their own data collection application, KoboCollect, which is essentially a clone of ODK Collect. KoboCollect is more visually compatible with Kobotools, but ODK Collect is generally more up-to-date and receives new features and bug-fixes at a faster rate than Kobotools.

Kobotoolbox also provides an online form builder, which is useful for building quick data collection forms without needing to write the XLSForm yourself. For more powerful forms, it is usually easier to author the form in XLSForm, which is fully compatible with the Kobotools server.

For further information visit the website at http://www.kobotoolbox.org/

Survey Solutions

Survey solutions is a suite of tools developed by the World Bank to support CAPI / CAWI / mobile data collection projects. It includes an online questionnaire designer, Android data collection application and “headquarters” server software, that enables project managers to oversee teams of enumerators.

Most of the tools are free to use, but the systems are not open source, so you will need to contact the Survey Solutions team to find out what's available for your project. While the World Bank resolve to offer the software for free, they may ask projects to pay for server costs, depending on the type of project.

For more information, see the Survey Solutions support portal: http://support.mysurvey.solutions/.

Options for Data and Document Storage

A Data and Document Store (DDS) is primarily for sharing project files within the team during the lifetime of the project. In this section we give a few common examples of tools that can be used for a DDS. See the separate document on DDS for more information.

OneDrive

OneDrive is a file-hosting service operated by Microsoft as part of its suite of online services. It allows users to store files in the cloud. Files can be synced to a PC and accessed from a web browser or a mobile device as well as shared publicly or with specific individuals.

OneDrive offers 5Gb of storage space free of charge; additional storage can be added either separately or through subscriptions to other Microsoft services including Office 365.

Like many syncing services, you can organise files within your OneDrive account into folders in the same way as you do in Windows on your local computer. Microsoft offers a syncing service for Windows and Mac, which lets you sync your OneDrive files to your local drive, allowing you to work with files offline. They also offer apps for iOS and Android to access your stored files.

On Windows 10, OneDrive offers selective syncing, which lets you view all your OneDrive files in Windows Explorer (as if they were stored locally), but to choose only certain files to have downloaded. This lets you view and access all your files without taking up space on your local drive. This is very useful if you have a large amount of content in OneDrive and a small disk in your local computer, but it's important to remember to download key files if you are going to be offline for an extended period.

For further information about OneDrive, visit the website at https://onerive.live.com/about/en-gb/

Dropbox

Dropbox is an easy to use and very popular file hosting service that gives access to files through the web. It enables easy sharing of files both as full access shares or through read-only links.

Like OneDrive, Dropbox offers Windows and Mac apps to sync your files to your local drive. The app creates a special folder on the user's computer and the contents of the folder are then synchronized to Dropbox's servers and to other computers and devices that the user has installed Dropbox on. Users can invite others to share one or more of their folders.

Dropbox Basic users are given 2Gb of free storage, but you can "earn" additional free space in a number of ways (see the web site for more details). Dropbox Plus and Dropbox Professional are paid subscriptions that both include 1Tb of space plus some additional features. Finally, there is Dropbox

Business which is geared towards organisations and groups. Pricing depends on the size of your team and your billing country.

For further information on Dropbox see the website at https://www.dropbox.com/ - there are also many introductory videos available on YouTube.

Google Drive

Google Drive is a file storage and synchronisation service. Google Drive offers 15Gb of free storage, with up to 30Tb offered through paid plans. Google Drive is a key component of G Suite, Google's monthly subscription offering for businesses and organisations.

As with the other cloud storage options, Google Drive offers Windows and Mac apps to sync your files to a folder on your local drive. The same app also allows you to backup any other folders on your local drive, but this also takes up space in your Google Drive account.

For further information, see the website at https://www.google.com/drive/ - again there are many introductory videos available on the web.

Google drive also links to Google's online office suite – Docs, Sheets and Slides. For information on these services, see https://www.google.com/docs/about/

Database Systems

Some projects' requirements for data storage and management can be addressed by careful management of data files, such as csv, Excel workbooks or json files. For project with more complex data structures, or projects that need to manage an evolving dataset over a period of time, you will likely require a database system to help manage your data.

The biggest decision to make is whether to run your database on your local computer or on a server.

- Local installations are easier to setup and have fewer dependencies. They also allow you to use your data whenever you need without having to worry about having a stable internet connection. However, only you have access to the database, and anyone else wanting access must take a copy, and changes made in different locations will not be automatically synchronised. You must also ensure you have adequate backups in place, in the same way as you would for regular files on your local disk.
- Databases running on a server are more complex to setup, and it is likely you will need the support of an IT / network expert to manage the system. Your organisation may already have infrastructure in place for one or more of the database types you require, so if you want to make use of a database on a server, talk to your IT team. They will also be able to advise on the best technical solutions for your project needs.

  Advantages of a database running on a remote server include:
  - It can act as a central location for your "truth" data, allowing your entire team to work with the same dataset.
  - All good systems have options for managing access levels, so you can choose what data to share with different users, and what permissions they have with those data.

- Remote servers generally have redundant hardware setups, so you are not relying on a single piece of hardware (e.g. your local hard drive).
- A good database setup will also have a way to record changes to your data, so you can maintain a unified record of all the changes that any member of your team makes during the course of the project.

The rest of this section lists different database packages that are available. It is not an exhaustive list, but covers many of the popular options.

MS Access

Microsoft Access is a powerful database management tool that allows a data manager to setup customised data entry forms to allow non-technical users to enter data and interact with the database effectively.

It is a good option if you require a database that can run locally, or shared among users on local network storage. Newer versions of MS Access also offer the option of creating cloud-based databases and applications, but these are generally less versatile than other solutions for cloud-based databases.

For more information, see the Microsoft site: https://products.office.com/en-gb/access

MySQL

MySQL is a free, open source SQL (structured query language) based database maintained by Oracle. It is one of the most flexible SQL database systems available and is easy to start using if you are new to structured databases.

Oracle provides installers for Windows, Mac and Linux to install MySQL locally: (https://dev.mysql.com/downloads/installer/), but the main way of using MySQL is to install it on a remote server that your project has access to.

MySQL is a good option to choose if you have a clear data structure with well-defined relationships between data objects and levels. Of the "SQL" based options, it is the easiest to get setup if you do not have much experience with database management.

PostgreSQL

PostgreSQL is another popular open source SQL-based language. It is similar to MySQL and shares many of the same traits, including the support for defined structures and data relationships. (https://www.postgresql.org/).

In some ways, it is far more powerful than MySQL, and is a good option to choose if you are working with extremely large structured datasets. It also has a powerful suite of tools for handling geographic objects and running location-based queries called PostGIS (https://postgis.net/). It is also considered more complex than MySQL and is not as easy to setup if you are unfamiliar with the options available.

MSSQL Server

MS-SQL is Microsoft's commercial SQL-based database. It is typically used in Windows Server / Dot-net environments and runs well in those situations. It is targeted more towards business users and teams already embedded into the Microsoft environment.

For more information, see the Microsoft website: https://www.microsoft.com/en-gb/sql-server/sql-server-2016

MongoDB

MongoDB is an open source noSQL database, developed and maintained by MongoDb Inc. (https://www.mongodb.com/). It is a popular database system for developers and data analysts, as it is highly scalable and performs well with extremely complex queries. Data are stored as JSON objects, which can be grouped into collections for easier management, and nested to account for multi-level datasets.

The database software itself is open source, but MongoDB Inc have a wide range of commercial offerings, including different levels of hosting and management.

For more information about the use of MongoDB, see their documentation site: https://docs.mongodb.com/getting-started/shell/tutorial/install-mongodb-on-windows/

CouchDB

CouchDB is another JSON-based noSQL database, maintained by the Apache group. (http://couchdb.apache.org/).

CouchDb provides a powerful API to interact with the database programmatically, and offers data replication options to allow users of applications to synchronise a dataset and continue to work offline.

Analysis Software

In this next section we look at a few of the more popular statistical analysis software packages.

R

R is a free, open source software environment for statistical computing and graphics. It runs on Windows, MacOS and UNIX platforms.

R provides a wide variety of statistical and graphical techniques and is highly extensible. One of its strengths is the ease with which well-designed publication quality plots can be produced.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License. R can easily be extended via "packages". There are about eight packages supplied with the R distribution and many more are available.

At the time of writing the latest version of R is version 3.4.3 which was released on 2017-11-30. It is a good idea to visit the website regularly for new releases.

For further details visit the website at https://www.r-project.org/

RStudio

RStudio is a front-end UI for running R. One of the challenges of using R is the basic nature of the interface and the challenges of needing to write commands for everything. RStudio is a popular option that gives UI controls for common tasks like importing datasets, loading package libraries and reviewing scripts.

For more information, see https://www.rstudio.com/products/RStudio/#Desktop

Stata

Stata is a general-purpose statistical software package. Its capabilities include data management, statistical analysis, graphics, simulations, regression and custom programming.

Stata licenses are generally for perpetual use although there is now the option to have an annual subscription. A single-user license can be installed on up to three of your personal computers as long as you are the sole user. See the website for pricing options.

At the time of writing, the latest version of Stata is version 15 (June 2017)

For further information about Stata visit the website at https://www.stata.com/

SPSS

SPSS (officially named IBM SPSS Statistics) is a widely used program for statistical analysis particularly in the social sciences. SPSS is reasonably easy to use either with the dropdown menus or via syntax files. SPSS is not cheap and there is an annual license fee. However, there is a GradPack available for current students which gives a substantial saving (up to 99%) on the standard annual license fee.

For further information about SPSS, visit the website at https://www.ibm.com/products/spss-statistics

GenStat

GenStat is a statistical software package with data analysis capabilities particularly in the field of agriculture. It is developed and marketed by VSN International Ltd (VSNi). GenStat licenses work on an annual subscription; you would need to contact VSNi directly for pricing information. Students can purchase the full version of GenStat at a reduced rate; the reduced rate is an annual fee and is only available to users in educational (degree granting) organisations.

For further information about GenStat, visit the website at https://www.vsni.co.uk/

Archiving Options / Repositories

In this section we briefly mention a couple of the more popular data and document repositories often used by researchers in the CGIAR.

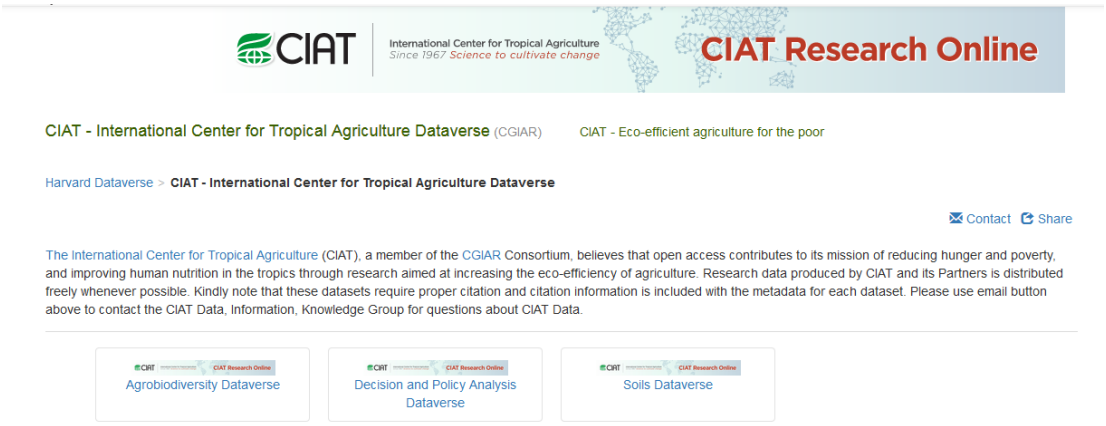
Dataverse

Dataverse is an open source web application for sharing, preserving, citing, exploring and analysing research data. It facilitates making data available to others and allows you to replicate others' work

more easily. Researchers, data authors, publishers, etc., all receive academic credit and web visibility.

A Dataverse repository hosts multiple dataverses which in turn contain datasets. Each dataset contains descriptive metadata and data files (including documentation and code to accompany the data). A separate document in this pack describes Dataverse in more detail.

CIAT and CCAFS both have their own dataverse and the image below shows the CIAT dataverse which contains three sub-dataverses: Agrobiodiversity, Decision and Policy Analysis, and Soils.

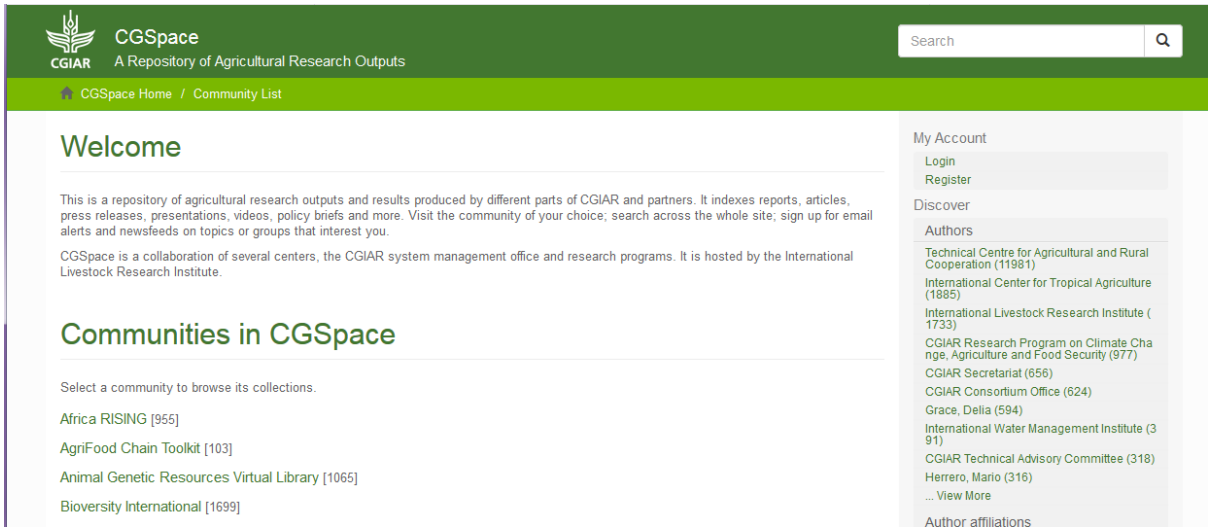


For more information about Dataverse, visit the website at https://dataverse.org/

DSpace

DSpace is software for building open digital repositories. It is free and easy to install and can be customised to fit the needs of any organisation. CIAT has its own Library Repository using DSpace which is available at http://ciat-library.ciat.cgiar.org:8080/jspui/

In addition, the CGIAR have set up CGSpace at https://cgspace.cgiar.org/ This is a repository of agricultural research outputs and results produced by CGIAR centres, initiatives and challenge programmes including CCAFS. The contents of the site can be browsed according to region, author, CGIAR centre, programme, etc.

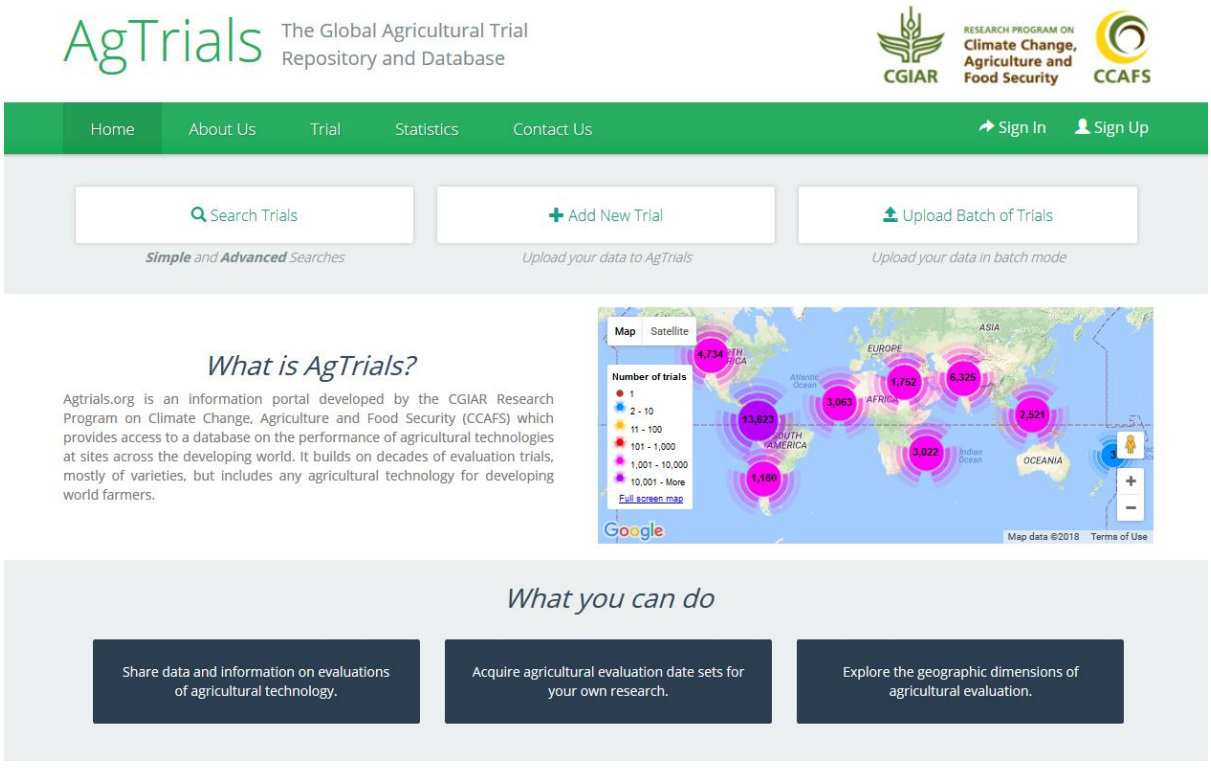


AgTrials

This is the Global Agricultural Trial Repository which is an information portal developed by CCAFS which provides access to a database of the performance of agricultural technologies at sites across the developing world. With the interface you can:

- Share data and information on evaluations of agricultural technology
- Acquire agricultural evaluation datasets for your own research
- Explore the geographical dimensions of agricultural evaluation.

The repository is available at http://agtrials.org/



CCAFS-Climate

The CCAFS-Climate data portal provides global and regional high-resolution climate datasets that serve as a basis for assessing the climate change impacts and adaptation in a variety of fields including biodiversity, agricultural and livestock production, and ecosystem services and hydrology.

The Climate portal is available at http://www.ccafs-climate.org/